Biophysical Journal Volume 106 April 2014 1638–1649

A Critical Evaluation of in silico Methods for Detection of Membrane Protein Intrinsic Disorder

Edward E. Pryor, Jr.[†] and Michael C. Wiener[†*]
[†]Department of Molecular Physiology and Biological Physics, Center for Membrane Biology, University of Virginia, Charlottesville, Virginia

ABSTRACT Intrinsically disordered regions in proteins possess important biological roles including transcriptional regulation, molecular recognition, and provision of sites for posttranslational modification. In three-dimensional crystallization of both soluble and membrane proteins, identification and removal of disordered regions is often necessary for obtaining crystals possessing sufficient long-range order for structure determination. Disordered regions can be identified experimentally, with techniques such as limited proteolysis coupled with mass spectrometry, or computationally, by using disorder prediction programs, of which many are available. Although these programs use various methods to predict disorder from a protein's primary sequence, they all were developed using information derived from soluble protein structures. Therefore, their performance and accuracy when applied to integral membrane proteins remained an open question. We evaluated the performance of 13 disorder prediction programs on a dataset containing 343 membrane proteins, and upon subdatasets containing only α-helical or β-barrel proteins. These programs were ranked using multiple metrics, including metrics specifically created for membrane proteins. Analysis of these data shows a clear distinction between programs that accurately predict disordered regions in membrane proteins and programs which perform poorly, and allows for the robust integration of in silico disorder prediction into our PSI:Biology membrane protein structural genomics pipeline.

INTRODUCTION

Integral membrane proteins are responsible for many cellular processes, and fall into functional classes that include ion channels, transporters, enzymes, and receptors, among others (1). In addition to their fundamental importance, membrane proteins are estimated to be targets for the majority of existing (and future) drugs (2). Therefore, determination of their structures is of both basic and practical significance. However, structure determination of membrane proteins is technically challenging (3), evidenced by the fact that membrane proteins comprise only ~2% of all the structures deposited in the Protein Data Bank (PDB) as of June 2013 (4). One of the many technical challenges in working with membrane proteins is an abundance of intrinsically disordered regions in these proteins. Approximately 70% of integral membrane proteins are predicted to contain disordered regions, compared to ~35% in soluble proteins (5). Disordered regions in proteins likely hinder crystallization efforts (6). In an examination of high throughput structure determination initiatives of membrane proteins, disordered regions in the target proteins lead to bottlenecks at each stage of the crystallography pipeline (7).

A long-standing dogma in biology is that the ordered three-dimensional structure of a protein is responsible for its function. However, numerous studies have shown that disordered regions in proteins are necessary for many functional roles in both soluble and membrane proteins (8–10). In soluble proteins, disordered regions are necessary for protein-nucleic acid interactions and transcriptional regulation (11), posttranslational modifications (12,13), and cell signaling (14). In membrane proteins, a notable example of a functionally significant disordered region is the disordered third intracellular loop (ICL3) of G-protein coupled receptors. The intrinsic disorder of ICL3 may enable it to interact with numerous protein targets and to induce multiple downstream signaling events (15). The disordered ICL3 has hindered crystallization of G-protein coupled receptors, often requiring the replacement of ICL3 by a stable domain such as T4 lysozyme (16–18). In another example, the C-terminal tail of voltage-activated potassium channels is intrinsically disordered and is hypothesized to interact with scaffolding proteins, such as the Post-Synaptic Density 95 protein (19). These disordered regions, in both soluble and membrane proteins, have been implicated in a number of human diseases, including cancer, cardiovascular disease, and diabetes (20).

A number of experimental approaches are available for the determination of disordered regions in proteins (for a detailed review, see Receveur-Bréchot et al. (21)), including hydrogen/deuterium exchange mass spectrometry (22), limited proteolysis (often in combination with mass spectrometry) (23–25), and NMR (26). However, these methods all require the expression and purification of a target protein, and are not generally practical for high throughput applications due to the added time and resources needed to perform these experiments. As a realistic alternative, many structural genomics groups

Submitted August 29, 2013, and accepted for publication February 25, 2014.

*Correspondence: mwiener@virginia.edu

Editor: Andreas Engel.

0006-3495/14/04/1638/12 $2.00

http://dx.doi.org/10.1016/j.bpj.2014.02.025

have incorporated in silico disorder prediction into the target selection and construct design phases of their pipelines. The Scottish Structural Proteomics Facility (http://www.sspf.ac.uk/) has incorporated the disorder predictor RONN (https://app.strubi.ox.ac.uk/RONN) into their target optimization tool TarO (27,28), whereas the Structural Proteomics in Europe consortium utilizes both FOLDINDEX (http://bip.weizmann.ac.il/fldbin/findex) and RONN to aid in their construct design (29). Disorder prediction with RONN has also been suggested as the first step in rational construct design for membrane proteins, specifically the truncation of disordered N- and C-termini (30). Lastly, the Protein Structure Initiative (PSI):Biology-funded New York Consortium on Membrane Protein Structure uses the program IUPRED (http://iupred.enzim.hu/) as part of their target selection pipeline (31).

Many different computational disorder prediction programs exist, and have been evaluated biannually at the Critical Assessment of Structure Prediction (CASP) experiments beginning with the CASP5 experiment in 2002 (32). To our knowledge, each available disorder prediction program was created and evaluated using information from soluble protein structures. Little information is available on the performance of disorder prediction programs on membrane proteins. One study, published by Xue et al. (33), examined the properties of disordered regions in a set of 120 membrane proteins. The results from that study showed that the amino-acid composition of disordered regions varied among α-helical, β-barrel, and soluble proteins. Additionally, that study characterized the performance, on a membrane protein dataset, of three disorder prediction programs in the PONDR family (Molecular Kinetics, Indianapolis, IN): VL3 (34), VLXT (35), and VSL2 (36). Missing from this study was an evaluation of how the many freely available disorder prediction programs perform on membrane proteins, specifically the prediction programs evaluated in the CASP experiments.

Our group, the Membrane Protein Structural Biology Consortium (MPSBC), is one of the nine specialized National Institutes of Health PSI:Biology centers devoted to membrane protein structure determination. As part of our pipeline, we have incorporated in silico disorder prediction to aid in the selection and design of membrane protein targets. Thus, we sought to address the open question of which prediction program performs best on membrane proteins. We used 13 disorder prediction programs to evaluate a nonredundant membrane protein structural database. Furthermore, each disorder predictor was tested on datasets comprised of only α-helical and β-barrel membrane proteins, and the results compared to the results from the full membrane protein dataset. Lastly, evaluation of these programs on the most recent CASP10 dataset allows for a comparison of performance between membrane proteins and soluble proteins. Using the results from this study, we have identified a prediction program that performs well on membrane proteins, PREDISORDER 1.1 (http://casp.rnet.missouri.edu/predisorder.html), and have incorporated this program into our MPSBC pipeline.

MATERIALS AND METHODS

Creation of the membrane protein dataset

PDB identifiers of membrane proteins were obtained from the Protein Data Bank of Transmembrane Proteins (PDBTM, http://pdbtm.enzim.hu/) (37). Specifically, the nonredundant portion of the dataset, containing sequences with <40% identity and a length longer than 30 residues, was utilized (accessed on June 7, 2013). This dataset was further curated to contain only structures determined by x-ray crystallography. We assigned regions as disordered using a method standard in the in silico disorder prediction community. Disordered regions were determined by aligning the sequence of the protein in the SEQRES records of the PDB files with the sequence derived from the Cα residues in the ATOM records of the PDB files. Any residue that was present in the records from the sequence of the protein used in the crystallization experiment, but not found in the ATOM records, was classified as a disordered residue. Membrane-spanning regions of each protein were determined by parsing the PDBTM XML file for each entry, which annotates transmembrane α-helices and β-strands. PDB identifiers of the 343 proteins found in the dataset used in this study are listed in Table S1 in the Supporting Material, and a file containing the membrane protein dataset is available from the authors upon request.

Selection of disorder prediction programs

Disorder prediction programs were chosen based on apparent popularity, as tabulated in Deng et al. (38), and also based upon their utilization in the CASP9 experiment (39). Additionally, the program was required to permit batch processing of the membrane protein dataset. In total, 13 disorder prediction programs were tested in this study, and are listed in Table 1. All programs were downloaded and installed on a local machine, with the exceptions of FOLDINDEX, which was run via a script written in Perl (http://www.perl.org/) that queried the FOLDINDEX webserver, and ESPRITZ 1.2 (http://biocomp.bio.unipd.it/espritz/), which was run by submitting a file containing each of the 343 sequences in FASTA format to the webserver. For the programs DISOPRED 2

TABLE 1 List of disorder predictors used in this study

Predictor	Website	Reference
DISEMBL (465)	http://dis.embl.de/	(41)
DISEMBL (coils)	http://dis.embl.de/	
DISEMBL (hot)	http://dis.embl.de/	
DISOPRED 2	http://bioinf.cs.ucl.ac.uk/disopred/	(42)
ESPRITZ 1.2	http://biocomp.bio.unipd.it/espritz/	(43)
FOLDINDEX	http://bip.weizmann.ac.il/fldbin/findex	(44)
GLOBPLOT	http://globplot.embl.de/	(45)
IUPRED (long)	http://iupred.enzim.hu/	(46)
IUPRED (short)	http://iupred.enzim.hu/	
PREDISORDER 1.1	http://casp.rnet.missouri.edu/predisorder.html	(47)
RONN	http://www.strubi.ox.ac.uk/RONN	(48)
SPINE-D	http://sparks.informatics.iupui.edu/SPINE-D/	(49)
VSL2B	http://www.dabi.temple.edu/disprot/predictorVSL2.php	(36)

(http://www.chemogenomix.com/chemogenomix/Disopred.html) and SPINE-D (http://sparks.informatics.iupui.edu/SPINE-D/), which require the creation of PSI-BLAST http://blast.ncbi.nlm.nih.gov/ profiles, the National Center for Biotechnology Information (NCBI, U.S. National Library of Medicine, National Institutes of Health, Bethesda, MD) nonredundant database, posted on June 8, 2013 and containing 26,236,801 sequences, was used. All methods were run using default parameters. It should be noted that the disorder predictors in the PONDR family (40) were not tested in this study due to the PONDR server only allowing a maximum of 50 predictions per user; however, the predictor VSL2B, developed by the same research group, was tested.

See also the literature (41–49).

Metrics to evaluate disorder prediction programs

A total of five metrics was utilized to evaluate each program, including three metrics created specifically to evaluate each program's performance on a membrane protein dataset. Because each disorder predictor is a binary classifier, assigning each residue in a sequence to be ordered or disordered, it is possible to calculate the number of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN) for each predicted sequence. A TP or TN occurs when a predictor correctly predicts an experimentally determined disordered or ordered residue. An FP occurs when a predictor incorrectly classifies an experimentally determined ordered residue as disordered, whereas an FN occurs when a predictor incorrectly classifies an experimentally determined disordered residue as ordered. From these four values, two metrics, balanced accuracy (ACC, Eq. 1) (50) and the Matthews correlation coefficient (MCC, Eq. 2) (51), can be calculated. These two metrics have previously been used to rank disorder prediction programs in previous CASP experiments (39,52–54). The value of ACC ranges from 0 (perfect inverse predictor) to 1 (perfect predictor), with 0.5 being a random predictor. The value of MCC ranges from −1 (perfect inverse predictor) to 1 (perfect predictor) with 0 being a random predictor:

$$ACC = \frac{\frac{TP}{(TP+FN)} + \frac{TN}{(TN+FP)}}{2}, \quad (1)$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{((TP+FP)(TP+FN)(TN+FP)(TN+FN))^{\frac{1}{2}}}. \quad (2)$$

A third metric calculates the percentage of predicted disorder in transmembrane regions (TM%). For each program, the number of disordered residues that were predicted within transmembrane regions was divided by the total number of predicted disordered residues. Analysis of the membrane protein dataset shows that no disordered residues exist in transmembrane regions; therefore, prediction programs with TM% values closer to 0% are posited to perform better on membrane proteins.

A disorder predictor that accurately predicts correct regions of disorder may incorrectly predict the number of disordered residues; however, predicting these regions correctly should be rewarded. For example, a disorder predictor may predict 10 residues on the N-terminus of a 100-residue protein to be disordered, when experimentally, 30 N-terminal residues are disordered. In this example, the balanced accuracy would be 0.563, which would be considered a poor prediction even though the predictor correctly identified the N-terminus as disordered. Calculation of the percentage of disordered regions found (R_D), allows for the analysis of how well each disorder predictor correctly identifies disordered regions. A disordered region in a structure is considered to be predicted by a specific program if at least one residue of that region is identified as disordered by that program. The dataset contains a total of 667 disordered regions.

The final metric compares the distribution of disordered region length in the membrane protein dataset with the distribution of disordered region lengths obtained from each predictor. An accurate predictor should have a distribution that closely resembles that of the membrane protein dataset. The χ^2 metric ($R\chi^2$) comparing two distributions equals 1 if the distributions are identical. Formally, this metric is defined by Eq. 3, where Predicted_i is the number of predicted disordered regions of length i, and PDB_i is the number of disordered regions of length i found in the membrane protein dataset. To account for the different number of disordered regions found in the membrane protein dataset and predicted by each program, the number of disordered regions of each length i is normalized by the total number of disordered regions found:

$$R\chi^2 = \sum_{i=1}^{\#\text{lengths}} \frac{\left(\left(\frac{\text{Predicted}_i}{\sum \text{Predicted}}\right) - \left(\frac{\text{PDB}_i}{\sum \text{PDB}}\right)\right)^2}{\left(\frac{\text{PDB}_i}{\sum \text{PDB}}\right)}. \quad (3)$$

Statistical significance of the metrics was calculated as previously described in the CASP7 experiment (53). Briefly, a bootstrapping method (55) was utilized where 80% of the targets were randomly selected 1000 times and the standard errors of the scores were calculated.

Receiver operating characteristic (ROC) curves (56) were also calculated, but not used in the final ranking of programs. Most disorder prediction programs output a probability of disorder for each residue; if the probability is above a certain value (typically 0.5), the residue is classified as disordered. A ROC curve is a plot of sensitivity ($TP/(TP+FN)$) against the false-positive rate ($FP/(TN+FP)$) for a number of probability threshold cutoffs, which were varied from 0 to 1 in increments of 0.001. For each cutoff value, residues with a probability above the cutoff were classified as disordered, and sensitivity and false-positive rate calculated. A ROC curve can be calculated for a disorder prediction program if it outputs an individual probability of disorder for each residue. Output from the programs GLOBPLOT (http://globplot.embl.de/), FOLDINDEX, DISEMBL (hot), DISEMBL (coils), and DISEMBL (465) (http://dis.embl.de/) do not contain individual probabilities for each residue, thus prohibiting calculation of ROC curves for these programs. Calculating the area under the ROC curve (AUC) is a measure of program performance. AUC values range from 0 to 1, with a value of 1 indicating a perfect predictor. AUC values for each ROC curve were calculated using SIGMAPLOT 8.02 (SPSS, IBM Corporation, Armonk, NY).

Calculation of crystal contacts

The Crystal Contact Analysis server (CryCo, http://ligin.weizmann.ac.il/cryco) (57) was utilized to determine the residues of each protein in the membrane protein dataset existing at crystal contacts. Using a Perl script, each protein was passed to the server (using the default settings), and the output parsed. Of the 343 proteins in the membrane protein dataset, crystal contact information for 316 was extracted. Crystal contact information for the remaining 27 proteins could not be calculated due to the large size of the PDB files.

Performance on CASP10 dataset

The 94 sequences used to evaluate disorder prediction programs in the latest CASP10 experiment were downloaded from the CASP website (http://predictioncenter.org/casp10/). Residues marked as *X* or *N* were not considered in the analysis. The CASP10 dataset is comprised of 24,190 residues, of which 1502 are disordered (6.2% disordered). Each of the 13 disorder prediction programs evaluated in this study was run on the CASP10 dataset, as described above, using the default parameters for each program.

RESULTS AND DISCUSSION

Performance of disorder prediction programs on a membrane protein dataset

The performance of 13 disorder prediction programs (listed in Table 1) was evaluated on a dataset comprising 343 membrane proteins whose structures were determined by x-ray crystallography. Of these proteins, 286 are α-helical, and 57 are β-barrel. In total, these proteins are comprised of 112,814 ordered residues and 9,591 disordered residues (8.5% disordered). The average resolution and length of the proteins in the dataset is 2.98 Å and 329 residues, respectively (Fig. 1, *A* and *B*). There is a total of 667 disordered regions in the dataset, with an average length of 14 residues. The distribution of disordered region lengths is shown in Fig. 1 *C*. Each of the 13 disorder prediction programs was run on the membrane protein dataset, and evaluated and ranked individually based upon five different metrics. The ranking of each individual metric was then summed to determine an overall ranking of predictors (Table 2). Full performance data of each of the programs are shown in Table S2 and intermetric correlation coefficients are listed in Table S3.

Balanced accuracy (ACC) and Matthews correlation coefficient (MCC), previously used in CASP evaluations (39,52–54), are strongly correlated with one another ($r = 0.846$, see Table S3). A clear separation exists between the two programs, *S*PINE-D and PREDISORDER 1.1, possessing highest ACC values (whose ACC values are statistically similar), and the next best performing program, DISOPRED 2, whose ACC value differs by 3.5% from that of PREDISORDER 1.1. Similar trends exist when examining the rankings based upon MCC, where the two predictors, DISOPRED 2 and ESPRITZ 1.2, possess the highest (and statistically similar) MCC values, whereas the MCC value for the next best performing predictor, PREDISORDER 1.1, differs by 13.1% from ESPRITZ 1.2. As observed with the best performing predictors, the three worst performing disorder predictors on the membrane protein dataset GLOBPLOT, IUPRED (long), and FOLDINDEX have similar ACC and MCC values, which are close to those of a random predictor (0.5 for ACC, and 0 for MCC).

We created three additional metrics to evaluate each of the disorder prediction programs. A common property often examined in the analysis of disorder prediction programs is the distribution of disordered region lengths in the dataset (shown in Fig. 1 *C* for the membrane protein dataset used in this study, and also previously shown for the CASP7–9 datasets (39,52,53)). The first of these new metrics, $R\chi^2$, examines the distributions of disordered region lengths between what is observed in the membrane protein dataset and what is predicted by each program, by calculating χ^2 between the two distributions. As an example, Fig. 2 *A* illustrates the normalized distribution of disordered region lengths for the membrane protein dataset (*black line*) overlaid with the most similar (PREDISORDER 1.1, *red line*) and least similar (DISEMBL (hot), *cyan line*) distributions from the disorder prediction programs. Plots comparing disordered region length distributions for each of the programs are shown in Fig. S1 in the Supporting Material.

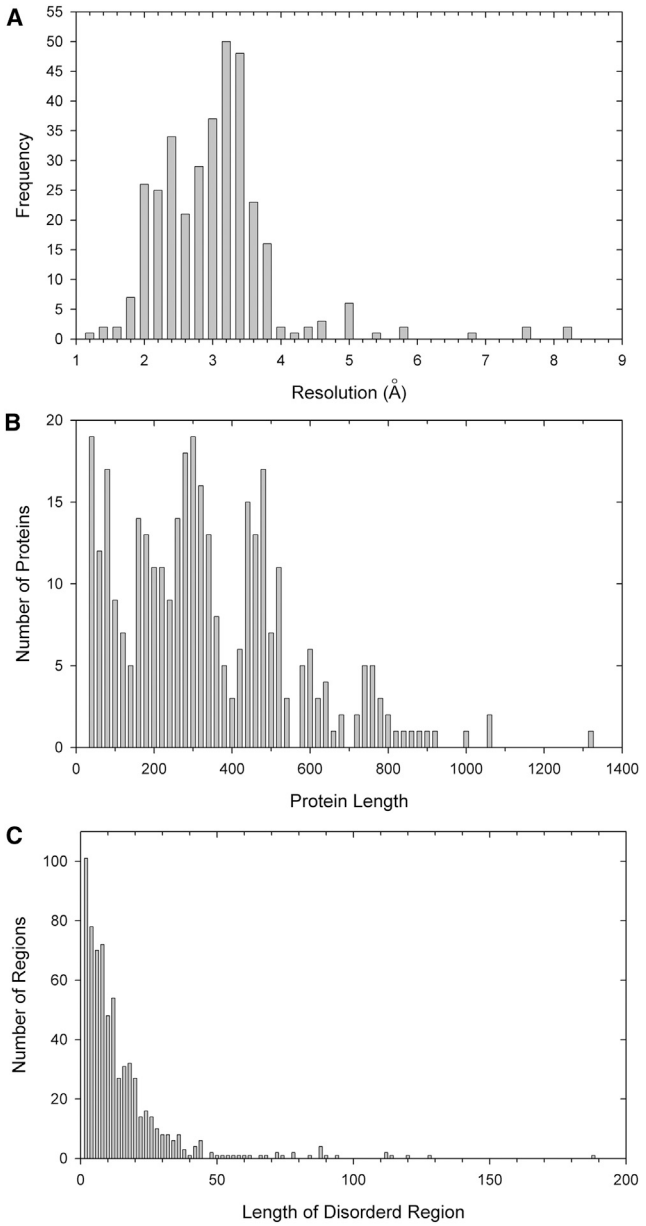

FIGURE 1 Properties of the 343 proteins in the membrane protein dataset. (*A*) Distribution of resolutions of proteins in the membrane protein dataset. Resolutions range from 1.20 Å (PDB:3M71) to 8.20 Å (PDB:4AC5) with an average resolution of 2.98 Å. (*B*) Distribution of protein lengths. Lengths of proteins in the membrane protein data set range from 31 residues (4FE1_M, 4H1W_B, and 1Q90_N) to 1306 residues (4F4C_A) with an average length of 329 residues. (*C*) Distribution of disordered region length in the membrane protein dataset. In total, 667 disordered regions exist with an average length of 14 residues.

TABLE 2 Performance of disorder predictors on the membrane protein dataset

Predictor	*ACC* ± SE	*MCC* ± SE	$R\chi^2$ ± SE	R_D ± SE (%)	*TM%* ± SE (%)	Ranking
PREDISORDER 1.1	0.785 ± 0.006 (1)	0.411 ± 0.007 (2)	20.618 ± 3.120 (1)	87.41 ± 0.63 (1)	3.89 ± 0.21 (3)	1 (8)
DISOPRED 2	0.757 ± 0.006 (2)	0.482 ± 0.009 (1)	23.108 ± 3.837 (1)	71.66 ± 0.91 (3)	2.40 ± 0.23 (2)	2 (9)
SPINE-D	0.790 ± 0.006 (1)	0.406 ± 0.008 (2)	24.764 ± 3.992 (1)	83.06 ± 0.81 (2)	14.44 ± 0.51 (7)	3 (13)
ESPRITZ 1.2	0.724 ± 0.006 (3)	0.474 ± 0.009 (1)	39.813 ± 6.184 (2)	68.52 ± 0.99 (4)	4.18 ± 0.32 (3)	3 (13)
VSL2B	0.738 ± 0.006 (3)	0.355 ± 0.008 (3)	24.790 ± 3.533 (1)	82.76 ± 0.76 (2)	4.72 ± 0.28 (4)	3 (13)
IUPRED (short)	0.669 ± 0.005 (5)	0.390 ± 0.010 (2)	61.804 ± 5.055 (3)	65.67 ± 0.94 (5)	3.99 ± 0.37 (3)	4 (18)
RONN	0.687 ± 0.005 (4)	0.289 ± 0.009 (4)	36.282 ± 7.814 (1)	49.18 ± 1.03 (6)	8.47 ± 0.40 (5)	5 (20)
DISEMBL (hot)	0.624 ± 0.004 (6)	0.372 ± 0.009 (3)	200.694 ± 15.380 (6)	32.23 ± 0.88 (8)	1.10 ± 0.17 (1)	6 (24)
DISEMBL (465)	0.657 ± 0.005 (5)	0.254 ± 0.009 (5)	130.601 ± 10.978 (5)	50.53 ± 0.98 (6)	7.44 ± 0.43 (5)	7 (26)
DISEMBL (coils)	0.616 ± 0.005 (6)	0.129 ± 0.006 (7)	95.008 ± 7.516 (4)	73.61 ± 0.88 (3)	16.01 ± 0.42 (8)	8 (28)
GLOBPLOT	0.580 ± 0.004 (7)	0.136 ± 0.008 (7)	77.002 ± 6.687 (3)	42.28 ± 0.93 (7)	10.86 ± 0.41 (6)	9 (30)
IUPRED (long)	0.588 ± 0.005 (7)	0.227 ± 0.011 (6)	128.536 ± 10.747 (5)	29.24 ± 1.01 (9)	5.34 ± 0.56 (4)	10 (31)
FOLDINDEX	0.573 ± 0.006 (7)	0.127 ± 0.011 (7)	116.418 ± 20.047 (4)	20.84 ± 0.98 (10)	20.33 ± 0.94 (9)	11 (37)

Ranking of each predictor for each metric is given in parentheses. The overall ranking of the programs was determined by summing each of the individual rankings for each evaluation metric. These results are given in the column "Ranking," with the sum of individual metric rankings given in parentheses.

Examining this data, there are small increases, not statistically significant, in $R\chi^2$ values of the first five programs. The first statistically significant difference in $R\chi^2$ occurs between RONN and ESPRITZ 1.2, with an increase of ~2% between the two programs. The rankings based upon the $R\chi^2$ metric correlate well with the commonly used ACC and MCC metrics, with correlation coefficients of 0.819, and 0.643, respectively (see Table S3).

An interesting observation is that >31% of the disordered regions in the membrane protein dataset have a length between one and five residues. This observation explains why certain disorder prediction programs, such as GLOBPLOT, FOLDINDEX, and the three programs in the DISEMBL suite, perform poorly according to this metric. Each of these programs uses a peak-width parameter, which defines the minimum length required to predict a disordered region. For example, the default peak width for the three DISEMBL programs is eight, meaning that these programs will not return any disordered regions between one and seven residues. The other two programs that performed poorly according to this metric are IUPRED (short) and IUPRED (long), which predict a large number of short disordered regions. IUPRED (short) predicted ~60% of disordered regions with a length between one and five residues, whereas IUPRED (long) predicted ~75% of disordered regions in the same range.

The fourth metric used in this study examined how accurately a predictor was able to correctly identify a region of disorder. This metric, percentage of disordered regions found (R_D), defined a correctly predicted region of disorder if at least one residue in that region was predicted as disordered. This metric was designed as an attempt to move from a very specific per-residue analysis of evaluation, to a more broad analysis. Using this metric, we are able to assess how well each predictor correctly identified residues within an area of a protein as disordered, which most commonly is the N- or C-terminus of the protein. In the membrane protein dataset, there is a total of 667 disordered regions (Fig. 1 *C*). PREDISORDER 1.1 had the highest R_D value, with SPINE-D and VSL2B having the next highest, and statistically identical, values. Overall, a low correlation (r = 0.437) exists between R_D, and the total number of predicted disordered residues (see Fig. S2 *A*). For example, the program DISEMBL (coils), which had the fourth highest R_D value, predicted the most disordered residues (53,676) of any program. The rankings based upon this metric are highly correlated with both ACC and $R\chi^2$ (r = 0.830 and 0.764, respectively); however, there is only a moderate correlation with the MCC metric (r = 0.571).

The last metric that we used was created to specifically examine the performance of disorder prediction programs on membrane proteins. An examination of the membrane protein dataset shows that no disordered residues exist in transmembrane α-helices or β-sheets, i.e., every residue classified as being in a transmembrane domain is present in the electron density of the crystal structure. This observation agrees with the hypothesis that, due to the nature of the lipid bilayer, membrane-spanning domains are highly stable (58,59). Thus, we surmised that a disorder prediction program that performs well on membrane proteins will predict very few disordered residues in transmembrane domains. The metric TM% calculates the percentage of disordered regions predicted in transmembrane domains. Six of the disorder predictors have TM% values below 5%, whereas the worst performing program (FOLDINDEX) predicts ~20% of its disordered residues in transmembrane domains. A surprising result is a predictor that performed poorly in the other four metrics, DISEMBL (hot), performed the best in this particular metric, only predicting ~1% of disordered residues in transmembrane domains. This result can be potentially explained by the observation that DISEMBL (hot) is a very conservative predictor that predicts the least amount of disordered residues (4019, 3.6% of the total) of all the programs. A moderate correlation exists between the total number of disordered residues predicted by a program and TM% (r = 0.537, see Fig. S2 *B*). Comparing the

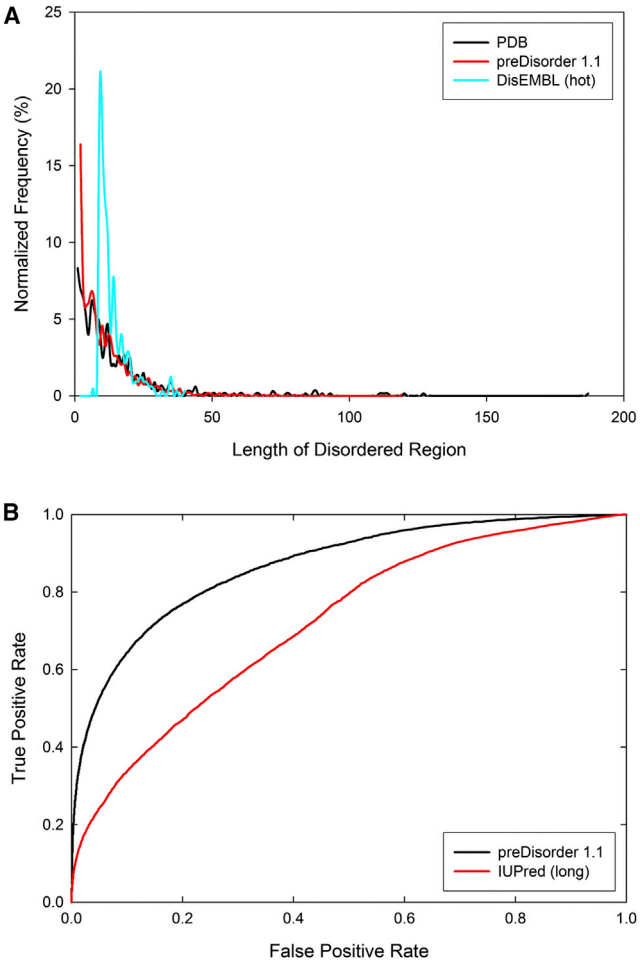

FIGURE 2 Examples of two metrics used to evaluate disorder prediction programs. (*A*) Plot illustrating χ^2 values between distributions of disordered region length ($R\chi^2$). Distributions of the lengths of disordered regions were calculated for the membrane protein dataset (*black line*) and each of the 13 disorder prediction programs. Shown in the plot are the distributions for the best (PREDISORDER 1.1, *red line*, $R\chi^2 = 20.618$) and worst (DISEMBL (hot), *cyan line*, $R\chi^2 = 200.694$) performing predictors determined by this metric. (*B*) ROC curves of best and worst performing predictors. The AUC for PREDISORDER 1.1 (*black line*) is 0.871, whereas the AUC for IUPRED (long) (*red line*) is 0.719.

rankings based upon TM% to the rankings based upon the two standard metrics of ACC and MCC, there is moderate correlation to ACC, and high correlation to MCC, with correlation coefficients of 0.418 and 0.731, respectively (see Table S3). However, when comparing the other two metrics created specifically for this study, $R\chi^2$ and R_D, there is little correlation between these metrics and the rankings based upon TM% ($r = 0.187$ and 0.170, respectively).

After evaluation of the membrane protein dataset, each predictor was ranked individually according to each metric. For each individual metric, programs received nonidentical ranks if a statistically significant difference ($p < 0.05$) existed between the values of that particular metric. The sum of the five ranks determined the final ranking of predictors (Table 2). Based upon these five metrics, the best performing disorder prediction program on the membrane protein dataset is PREDISORDER 1.1, which ranked in the top three for each of the five metrics used. To determine how the 13 prediction programs correlate to one another on the same protein target, individual ACC values for each protein in the membrane protein dataset were calculated for each prediction program and compared to the average ACC for that protein (see Fig. S3 and Table S4). This allows for the identification of proteins in which all of the disorder prediction programs performed well or performed poorly. Representative examples of membrane proteins, with intrinsic disorder predicted well or poorly (according to average ACC values) by each of the disorder predictors, are shown in Fig. 3. Additionally, in the membrane protein dataset, disordered regions exist only in extramembranous regions, completely absent from transmembrane domains of the proteins. Fig. S4 shows representative examples of well-predicted extramembranous regions by the top-ranked disorder prediction program, PREDISORDER 1.1.

Another metric commonly calculated for disorder prediction programs is AUC, the area under the ROC curve. AUC values were not used in the final rankings, as they could not be calculated for all of the programs. AUC values for eight of the disorder prediction programs are shown in Table 3, and ROC curves for the best (PREDISORDER 1.1) and worst (IUPRED (long)) performing programs, according to AUC, are shown in Fig. 2 *B*. Graphs of the ROC curves for each of the eight programs are shown in Fig. S5. An examination of these values shows a strong agreement between the final rankings shown in Table 1, and the rankings based upon AUC, with PREDISORDER 1.1 still performing the best on the membrane protein dataset.

Effects of crystal contacts on protein disorder

We sought to determine whether residues in the membrane protein dataset, which are predicted in silico to be disordered, become ordered when at the site of a crystal contact. In other words, do crystal contacts induce structure in normally disordered regions of the protein? To address this question, we examined the occurrence of FPs in the membrane protein dataset. False-positives exist when a prediction program classifies a residue as disordered, but the residue is ordered in the protein structure. From this, we can calculate the false-positive rate ($FPR = FP/(FP + TN)$) for each prediction program. Using the CryCo server (57) residues existing at crystal contacts for 316 proteins in the membrane protein dataset were extracted. Overall, for each of the prediction programs, between 11 and 20% of the false positives exist at a crystal contact (see Table S5). Additionally, an FPR for residues existing at crystal contacts was calculated by determining the number of FPs and TNs that exist at crystal contacts. Compared to the overall FPR, the FPR for residues existing at crystal contacts is statistically higher ($p < 0.05$) for 11 of the 13 programs (see

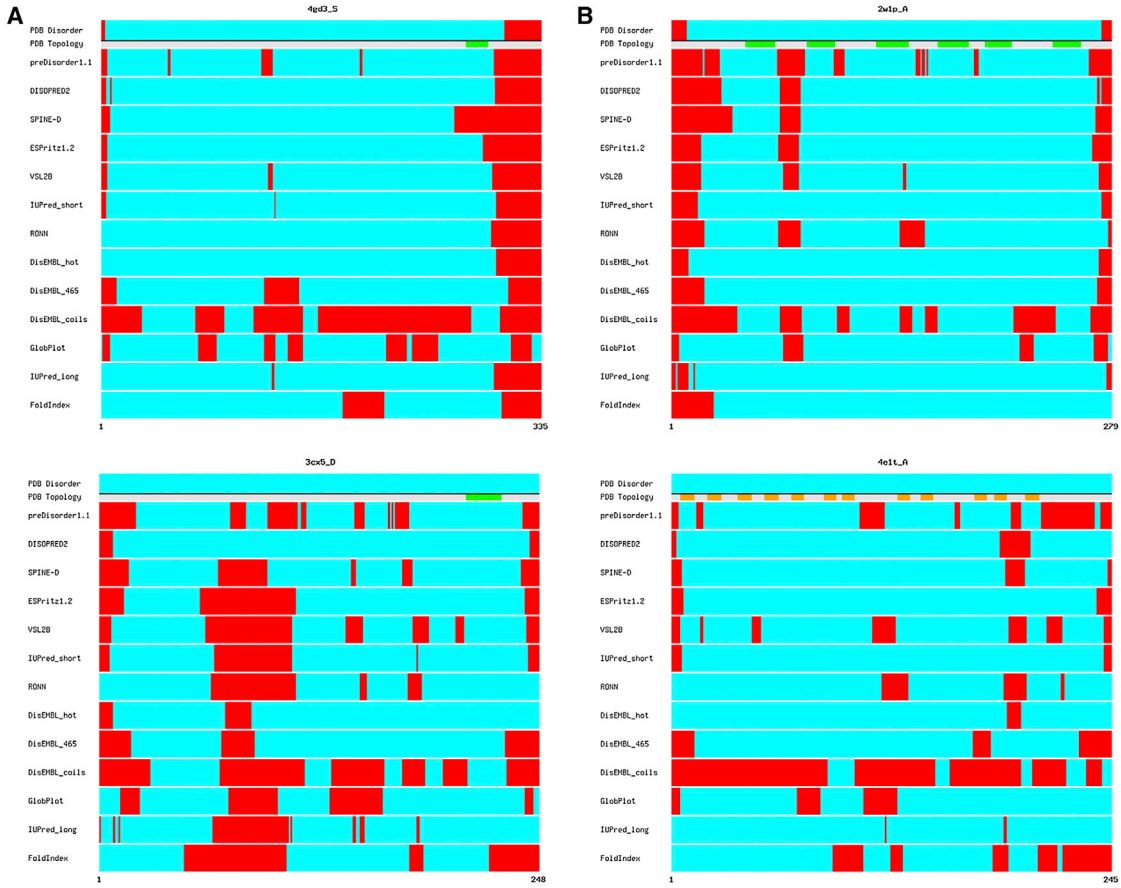

FIGURE 3 Representative examples of disorder prediction by all 13 programs. (*Red*) Disordered residues; (*blue*) ordered residues; (*green*) PDB topology for α-helices; (*orange*) β-strands. (*A*) Examples of disorder prediction on monotopic membrane proteins where each of the 13 prediction programs performed well (*top*, average $ACC = 0.903$) and poorly (*bottom*, average $ACC = 0.363$). (*B*) Examples of disorder prediction on polytopic membrane proteins where each of the 13 prediction programs performed well (*top*, average $ACC = 0.889$) and poorly (*bottom*, average $ACC = 0.411$).

Table S5), with the two outliers, IUPRED (long) and FOLDINDEX, being the worst performing predictors (Table 2). These results suggest, intriguingly, that crystal contacts may play a role in ordering intrinsically disordered extramembranous regions of integral membrane proteins.

Performance comparison on α-helical and β-barrel proteins

To determine whether differences in prediction performance exist between α-helical and β-barrel membrane proteins, the membrane protein dataset was divided into these two classes:

1. The α-helical dataset comprises 286 proteins, for a total of 87,311 residues, 7908 of which are disordered (9.1% disordered). Additionally, there are 544 disordered regions with an average disordered region length of 15 residues.
2. The β-barrel dataset, consisting of 57 proteins, contains a total of 25,503 residues, 1683 of which are disordered (6.6% disordered). There are 123 disordered regions found in the proteins in this dataset, with an average disordered region length of 14 residues.

TABLE 3 Area under the curve (AUC) measurements for selected disorder predictors

Predictor	*AUC*
PREDISORDER 1.1	0.871
SPINE-D	0.868
DISOPRED 2	0.864
ESPRITZ 1.2	0.843
VSL2B	0.824
IUPRED (short)	0.799
RONN	0.761
IUPRED (long)	0.719

In comparison to the full membrane protein dataset, the α-helical dataset is slightly more disordered whereas the β-barrel dataset is slightly less disordered. Both the α-helical and β-barrel datasets have an average disordered region length that is similar to the full membrane protein dataset.

The five metrics used to evaluate the full membrane protein dataset were used to evaluate the α-helical and β-barrel datasets (Fig. 4). The performance data for both datasets are

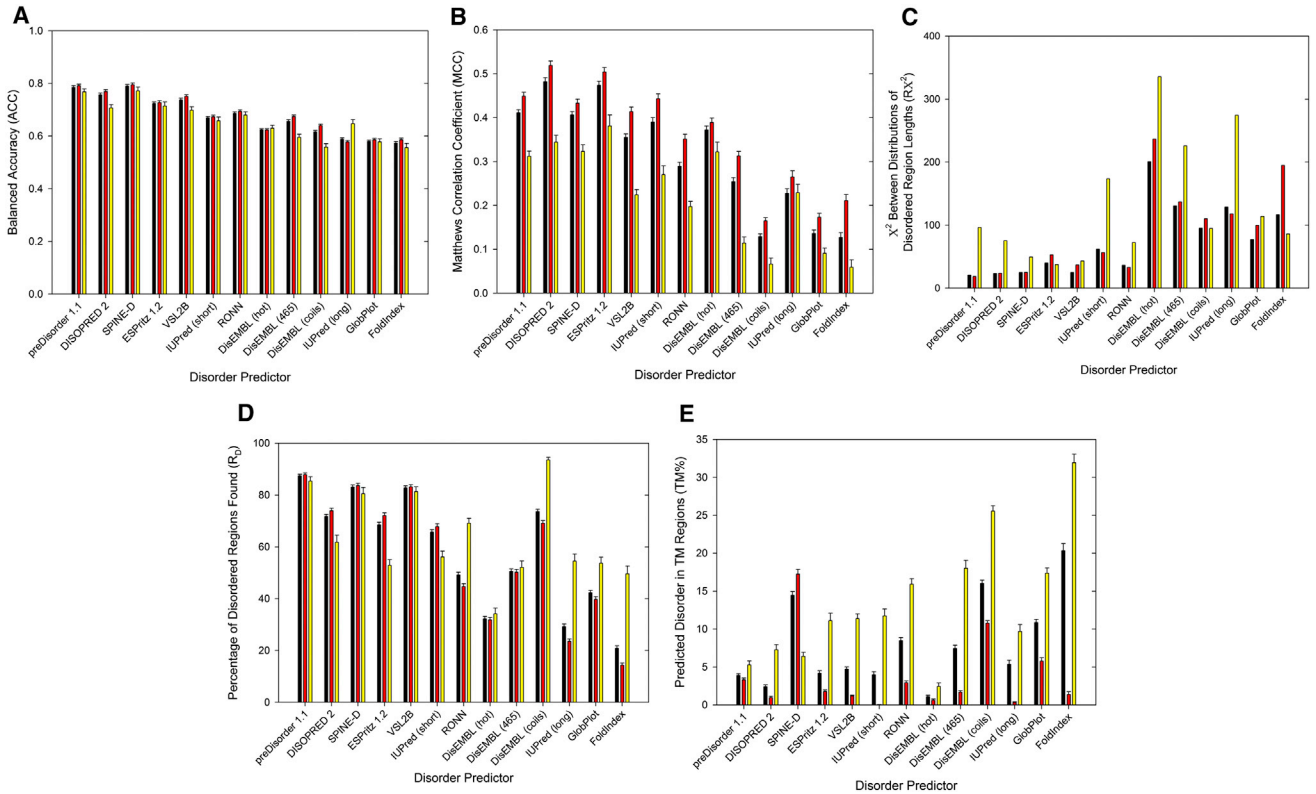

FIGURE 4 Performance of disorder prediction programs on α-helical and β-barrel proteins. (*Black*) Performance on all 343 proteins in the membrane protein dataset; (*red*) performance on the 286 α-helical proteins; (*yellow*) performance on the 57 β-barrel proteins. Results for each of the five metrics used to evaluate each disorder prediction program are shown. (*A*) Balanced accuracy (*ACC*). (*B*) Matthews correlation coefficient (*MCC*). (*C*) χ^2 comparing distributions of disordered region lengths ($R\chi^2$). (*D*) Percentage of disordered regions found (R_D). (*E*) Predicted disorder in transmembrane regions (TM%).

shown in Table S6 and Table S7. Examination of the ACC metric shows that, overall, only slight differences exist between each protein class, and performance is similar on each (Fig. 4 *A*). Of the 13 disorder prediction programs, performance on the α-helical dataset is slightly better than on the β-barrel dataset for 11 of the programs. The largest deviation exists with the program DISEMBL (coils), where the ACC value (from the first metric) for the α-helical dataset is 10.3% greater than that of the β-barrel dataset. The two programs where performance on the β-barrel dataset is better than their performance on the α-helical dataset are DISEMBL (hot) and IUPRED (long), in which the ACC values for the β-barrel dataset are 0.9 and 8.7% higher than the α-helical dataset, respectively. Contrary to what is observed with the ACC metric, there are large differences between each topology class when the MCC metric (the second metric) is used (Fig. 4 *B*). MCC values were greater for the α-helical dataset than on the β-barrel dataset for each predictor examined, and range from 6.9% (IUPRED (long)) to 38.3% (DISEMBL (465)) larger.

Examination of $R\chi^2$, the third metric, shows a similar trend to that observed with the MCC metric. For 10 of the disorder predictors, performance on α-helical proteins was better than on β-barrel proteins (Fig. 4 *C*). The largest discrepancy in performance is seen with IUPRED (long), where $R\chi^2$ for the α-helical dataset is 46.7% lower than the value for the β-barrel dataset. The only instance where performance on the β-barrel dataset is significantly better than the α-helical dataset is seen with the program FOLDINDEX, where there is a 32.5% difference between the two values. These results indicate that even though the average length of disordered region is the same for both topology datasets, the distributions of the disordered region lengths differ in each dataset.

The fourth metric examined, R_D, shows varied performance between each prediction program (Fig. 4 *D*). For five of the disorder prediction programs, there is no significant difference in R_D value between each transmembrane secondary structure dataset (<4% difference); however, three of the predictors perform better on the α-helical dataset, whereas five of the predictors perform better on the β-barrel dataset. Of the three programs that perform better on the α-helical dataset, the largest difference is seen with ESPRITZ 1.2, whose R_D is 19.2% higher for the α-helical dataset than for the β-barrel dataset. In contrast, four of the five programs that perform better on the β-barrel dataset (RONN, DISEMBL (coils), IUPRED (long), and FOLDINDEX) have R_D values that are >24% higher than the R_D value for

the α-helical dataset; the largest difference is observed with the program FOLDINDEX (35.3% greater).

The fifth (and last) metric used to evaluate the two topology datasets, TM%, is shown in Fig. 4 *E*. Two observations emerge from examination of this metric: The first observation is that, with one exception (SPINE-D), all of the disorder predictors perform better on the α-helical dataset; the second observation is that the performance on the α-helical dataset by these programs is quite good, with eight of the 13 programs having TM% values under 2%. There are some striking performance differences by some of the programs on each topology dataset. For example, the program FOLDINDEX has a TM% value of 1.38% for the α-helical dataset and a TM% value of 31.92% for the β-barrel dataset. The observed variations in performance between the two topology datasets might have to do with the inherent differences between α-helical and β-barrel proteins, specifically the number of membrane-spanning domains. The proteins in the β-barrel dataset have an average of 15 transmembrane domains, whereas the proteins in the α-helical dataset have an average of five transmembrane domains. Perhaps the decreased performance on the β-barrel dataset is due to more transmembrane domains being present in these proteins, increasing the probability that a predicted disordered residue will reside in that domain.

Similar to the analysis performed with the results from the full membrane protein dataset, for both transmembrane secondary structure datasets each predictor was ranked based upon its performance using the five metrics. The rankings for the α-helical dataset are shown in Table 4, and the rankings for the β-barrel dataset are shown in Table 5. In addition to ranking the predictors based upon performance of each transmembrane secondary structure dataset, these results also allow users to examine the contributions of each transmembrane secondary structure class on the full membrane protein dataset. Overall, these rankings differ slightly from the rankings calculated based on the full membrane protein dataset shown in Table 2; however, the same programs rank in the top five on all three datasets. The best performing programs for α-helical proteins is PREDISORDER 1.1 and DISOPRED 2 (tied), whereas SPINE-D performs the best on β-barrel proteins. The best performing program on the full membrane protein dataset, PREDISORDER 1.1, ranks second for the β-barrel dataset.

TABLE 4 Rankings of disorder predictors on α-helical proteins

Predictor	*ACC*	*MCC*	$R\chi^2$	R_D	*TM%*	Ranking
PREDISORDER 1.1	1	2	1	1	5	1 (10)
DISOPRED 2	2	1	1	3	3	1 (10)
VSL2B	2	3	2	2	3	2 (12)
ESPRITZ 1.2	3	1	2	3	4	3 (13)
SPINE-D	1	2	1	2	8	4 (14)
IUPRED (short)	5	2	3	4	1	5 (15)
RONN	4	4	1	6	5	6 (20)
DISEMBL (465)	5	5	5	5	4	7 (24)
DISEMBL (hot)	7	3	6	8	2	8 (26)
DISEMBL (coils)	6	8	4	4	7	9 (29)
IUPRED (long)	8	6	4	9	2	9 (29)
FOLDINDEX	8	7	5	10	3	10 (33)
GLOBPLOT	8	8	4	7	6	10 (33)

Ranking of each predictor for each metric is given in parentheses. The overall ranking of the programs was determined by summing each of the individual rankings for each evaluation metric. These results are given in the column "Ranking," with the sum of individual metric rankings given in parentheses.

TABLE 5 Rankings of disorder predictors on β-barrel proteins

Predictor	*ACC*	*MCC*	$R\chi^2$	R_D	*TM%*	Ranking
SPINE-D	1	1	1	2	2	1 (7)
PREDISORDER 1.1	1	2	2	2	2	2 (9)
DISOPRED 2	2	1	1	4	3	3 (11)
VSL2B	2	3	1	2	4	4 (12)
ESPRITZ 1.2	2	1	1	5	4	5 (13)
RONN	2	3	1	3	5	6 (14)
DISEMBL (hot)	3	1	4	6	1	7 (15)
IUPRED (short)	3	2	3	4	4	8 (16)
IUPRED (long)	3	3	4	4	4	9 (18)
DISEMBL (coils)	5	5	2	1	6	10 (19)
DISEMBL (465)	4	4	3	5	5	11 (21)
GLOBPLOT	4	4	3	5	5	11 (21)
FOLDINDEX	5	5	1	5	7	12 (23)

Ranking of each predictor for each metric is given in parentheses. The overall ranking of the programs was determined by summing each of the individual rankings for each evaluation metric. These results are given in the column "Ranking," with the sum of individual metric rankings given in parentheses.

Comparison to soluble protein prediction

Lastly, we sought to assess whether the top performing disorder prediction programs perform differently on membrane proteins compared to soluble proteins. To test these differences, the 13 disorder prediction programs were run using the 94 protein sequences from the most recent CASP10 dataset (see Table S8). A comparison of ACC and MCC values for each of the prediction programs on the membrane protein and CASP10 datasets is shown in Fig. 5. In terms of ACC, performance on both datasets is consistent between both protein datasets (Fig. 5 *A*). The ratio of ACC (membrane protein dataset) to ACC (CASP10 dataset) ranges from 1.11 (DISEMBL (coils)) to 0.98 (DISEMBL (465)), with an average ratio of 1.03 $\pm$ 0.04. Examination of MCC values shows larger variations in performance between the membrane protein dataset and the CASP10 dataset for most of the disorder prediction programs (Fig. 5 *B*). The ratio of MCC (membrane protein dataset) to MCC (CASP10 dataset) ranges from 2.48 (DISEMBL (coils)) to 0.97 (GLOBPLOT), with an average ratio of 1.37 $\pm$ 0.39.

According to the MCC metric, performance of all but one disorder predictor (GLOBPLOT) is better on the membrane protein dataset, a striking result. This difference in performance was also observed when the PONDR family of

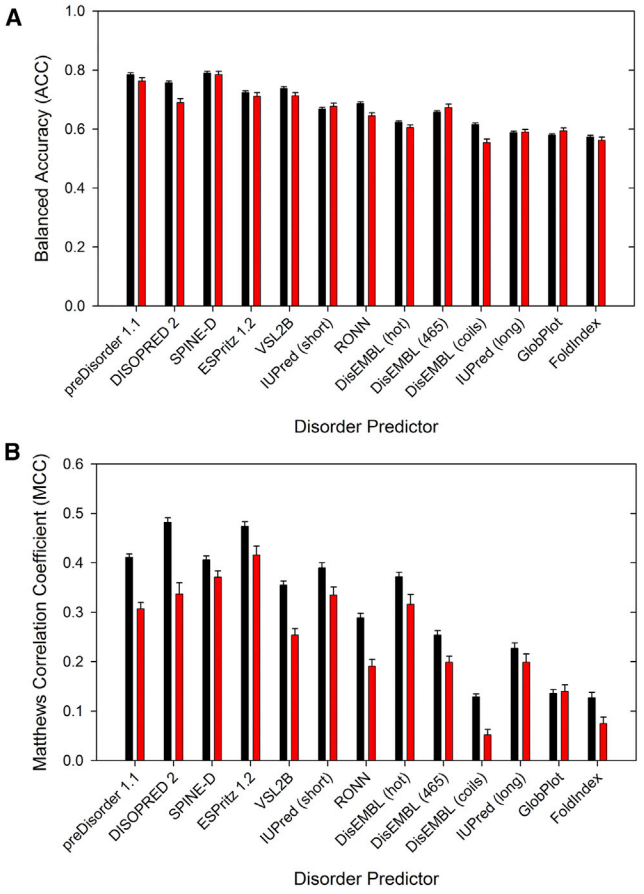

FIGURE 5 Comparison of membrane protein and CASP10 datasets. The 94 sequences in the CASP10 dataset were analyzed with each of the 13 disorder predictors tested in this study. (*Black*) Values for the membrane protein dataset; (*red*) values for the CASP10 dataset. (*A*) Comparison of balanced accuracy (*ACC*) between each of the predictors shows that very few differences exist between the two datasets. (*B*) Comparison of the Matthews correlation coefficient (*MCC*) between each disorder predictor shows larger performance discrepancies between the two datasets with performance generally being better on the membrane protein dataset.

disorder predictors was evaluated using a membrane protein dataset (33). In that study, the better performance on a membrane protein dataset was observed with the AUC metric. The better performance on the membrane protein dataset is particularly interesting, considering that these programs were created using knowledge of soluble protein structures (not membrane protein structures). We also evaluated the CASP10 dataset using the two metrics created specifically for this study, $R\chi^2$ and R_D (see Fig. S6). Using the $R\chi^2$ metric, there are few observable differences between the top performing programs on both datasets, whereas analysis with R_D shows slightly better performance on the membrane protein dataset for the top performing programs.

CONCLUSIONS

The main goal of this study was to evaluate the performance of a number of available disorder prediction programs on a dataset comprised of integral membrane proteins. Because these programs were created and tested based upon knowledge derived from soluble protein structures, we wanted to examine the performance of these programs when applied to membrane proteins. On a dataset comprised of both α-helical and β-barrel proteins, the program PREDISORDER 1.1 was the overall best performing, and also ranked first in the α-helical dataset and second in the β-barrel dataset. Additionally, there are no major differences in performance of these programs between membrane proteins and soluble proteins, although we obtained the surprising result that performance on membrane proteins is slightly better than on soluble proteins.

Overall, these results indicate that some of the disorder prediction programs tested here are suitable for use on integral membrane proteins. Although our performing this comparative analysis was motivated by our use of in silico disorder prediction in our x-ray crystallographic structure determination pipeline, other applications exist. For example, removal of disordered regions would likely reduce the size of the protein detergent complex, and smaller protein detergent complex size will generally improve the quality of solution NMR spectra. Also, researchers studying the biological importance of disordered regions in transmembrane proteins, which have an increasing number of functional roles, can use a disorder prediction program to identify these regions.

Based upon our analysis, we have incorporated PREDISORDER 1.1 for use in the target selection and constructed design steps of our MPSBC crystallographic structure determination pipeline. Typically, a potential target protein is examined, and if large regions of disorder exist between predicted transmembrane helices, the target is given a low priority or not selected. Regions of disorder located at the termini of the protein are truncated if they are longer than 15 residues, fall outside of predicted transmembrane domains, and are not implicated to play significant functional roles. Other implementations of in silico disorder prediction programs are certainly also reasonable. For example, DISOPRED 2 could be used on predicted α-helical membrane proteins, and SPINE-D could be used on predicted β-barrel membrane proteins. (We note that we find no significantly noticeable difference in practical utility if we use DISOPRED 2 instead of PREDISORDER 1.1.) Alternatively, several programs (e.g., PREDISORDER 1.1 and DISOPRED 2, or PREDISORDER 1.1 and SPINE-D) can be used and compared; such an approach has been adapted by structural genomics efforts directed to soluble proteins (29).

We calculated an unweighted sum of each metric's numerical ranking for the final score of each program. Other potential users may wish to use different weighting schemes, so we have provided all of the individual values of each of the characteristic metrics. With increasing numbers of membrane protein structures available,

designers of disorder prediction programs can utilize this information in the creation and training of their programs. There may be additional value gained from integrating information from other sources such as hydropathy plots and transmembrane domain prediction, which may prove quite helpful in predicting disordered regions in integral membrane proteins. We encourage the disorder prediction community to incorporate a testing set of integral membrane proteins into future program development and evaluation.

SUPPORTING MATERIAL

Six figures and eight tables are available at http://www.biophysj.org/biophysj/supplemental/S0006-3495(14)00262-8.

The authors thank the members of the Membrane Protein Structural Biology Consortium, particularly Drs. Peter Horanyi, Mark Dumont, and Michael Malkowski, for useful discussion. Dr. David Cooper provided useful suggestions on calculation of crystal contacts.

This work was supported by Protein Structure Initiative: Biology grant No. U54-GM094611 (National Institutes of Health) for membrane protein structural genomics.

REFERENCES

1. von Heijne, G. 2006. Membrane-protein topology. *Nat. Rev. Mol. Cell Biol.* 7:909–918.
2. Bakheet, T. M., and A. J. Doig. 2009. Properties and identification of human protein drug targets. *Bioinformatics.* 25:451–457.
3. Wiener, M. C. 2004. A pedestrian guide to membrane protein crystallization. *Methods.* 34:364–372.
4. Berman, H. M., J. Westbrook, …, P. E. Bourne. 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.
5. Yang, J. Y., M. Q. Yang, …, X. Huang. 2008. Investigation of transmembrane proteins using a computational approach. *BMC Genomics.* 9 (*Suppl 1*):S7.
6. Oldfield, C. J., B. Xue, …, V. N. Uversky. 2013. Utilization of protein intrinsic disorder knowledge in structural proteomics. *Biochim. Biophys. Acta.* 1834:487–498.
7. Johnson, D. E., B. Xue, …, V. N. Uversky. 2012. High-throughput characterization of intrinsic disorder in proteins from the Protein Structure Initiative. *J. Struct. Biol.* 180:201–215.
8. Dyson, H. J., and P. E. Wright. 2005. Intrinsically unstructured proteins and their functions. *Nat. Rev. Mol. Cell Biol.* 6:197–208.
9. Babu, M. M., R. W. Kriwacki, and R. V. Pappu. 2012. Structural biology. Versatility from protein disorder. *Science.* 337:1460–1461.
10. Fong, J. H., B. A. Shoemaker, and A. R. Panchenko. 2012. Intrinsic protein disorder in human pathways. *Mol. Biosyst.* 8:320–326.
11. Dyson, H. J. 2012. Roles of intrinsic disorder in protein-nucleic acid interactions. *Mol. Biosyst.* 8:97–104.
12. Iakoucheva, L. M., P. Radivojac, …, A. K. Dunker. 2004. The importance of intrinsic disorder for protein phosphorylation. *Nucleic Acids Res.* 32:1037–1049.
13. Gao, J., and D. Xu. 2012. Correlation between posttranslational modification and intrinsic disorder in protein. *Pac. Symp. Biocomput.* 17:94–103. *http://psb.stanford.edu/psb-online/proceedings/psb12/gao.pdf*.
14. Iakoucheva, L. M., C. J. Brown, …, A. K. Dunker. 2002. Intrinsic disorder in cell-signaling and cancer-associated proteins. *J. Mol. Biol.* 323:573–584.
15. Woods, A. S. 2010. The dopamine D_4 receptor, the ultimate disordered protein. *J. Recept. Signal Transduct. Res.* 30:331–336.
16. Cherezov, V., D. M. Rosenbaum, …, R. C. Stevens. 2007. High-resolution crystal structure of an engineered human β2-adrenergic G protein-coupled receptor. *Science.* 318:1258–1265.
17. Rosenbaum, D. M., V. Cherezov, …, B. K. Kobilka. 2007. GPCR engineering yields high-resolution structural insights into β2-adrenergic receptor function. *Science.* 318:1266–1273.
18. Rosenbaum, D. M., C. Zhang, …, B. K. Kobilka. 2011. Structure and function of an irreversible agonist-β_2 adrenoceptor complex. *Nature.* 469:236–240.
19. Magidovich, E., S. J. Fleishman, and O. Yifrach. 2006. Intrinsically disordered C-terminal segments of voltage-activated potassium channels: a possible fishing rod-like mechanism for channel binding to scaffold proteins. *Bioinformatics.* 22:1546–1550.
20. Uversky, V. N., C. J. Oldfield, …, A. K. Dunker. 2009. Unfoldomics of human diseases: linking protein intrinsic disorder with diseases. *BMC Genomics.* 10 (*Suppl 1*):S7.
21. Receveur-Bréchot, V., J. M. Bourhis, …, S. Longhi. 2006. Assessing protein disorder and induced folding. *Proteins.* 62:24–45.
22. Balasubramaniam, D., and E. A. Komives. 2013. Hydrogen-exchange mass spectrometry for the study of intrinsic disorder in proteins. *Biochim. Biophys. Acta.* 1834:1202–1209.
23. Fontana, A., P. Polverino de Laureto, …, M. Zambonin. 1997. Probing the partly folded states of proteins by limited proteolysis. *Fold. Des.* 2:R17–R26.
24. Hubbard, S. J. 1998. The structural aspects of limited proteolysis of native proteins. *Biochim. Biophys. Acta.* 1382:191–206.
25. Cohen, S. L., and B. T. Chait. 2001. Mass spectrometry as a tool for protein crystallography. *Annu. Rev. Biophys. Biomol. Struct.* 30:67–85.
26. Dyson, H. J., and P. E. Wright. 2002. Insights into the structure and dynamics of unfolded proteins from nuclear magnetic resonance. *Adv. Protein Chem.* 62:311–340.
27. Overton, I. M., and G. J. Barton. 2011. Computational approaches to selecting and optimizing targets for structural biology. *Methods.* 55:3–11.
28. Overton, I. M., C. A. van Niekerk, …, G. J. Barton. 2008. TARO: a target optimization system for structural biology. *Nucleic Acids Res.* 36 (Web Server issue):W190–W196.
29. Esnouf, R. M., R. Hamer, …, J. Prilusky. 2006. Honing the in silico toolkit for detecting protein disorder. *Acta Crystallogr. D Biol. Crystallogr.* 62:1260–1266.
30. Drew, D., and H. Kim. 2012. Screening for high-yielding *Saccharomyces cerevisiae* clones: using a green fluorescent protein fusion strategy in the production of membrane proteins. *Methods Mol. Biol.* 866:75–86.
31. Punta, M., J. Love, …, B. Rost. 2009. Structural genomics target selection for the New York consortium on membrane protein structure. *J. Struct. Funct. Genomics.* 10:255–268.
32. Melamud, E., and J. Moult. 2003. Evaluation of disorder predictions in CASP5. *Proteins.* 53 (*Suppl 6*):561–565.
33. Xue, B., L. Li, …, A. K. Dunker. 2009. Analysis of structured and intrinsically disordered regions of transmembrane proteins. *Mol. Biosyst.* 5:1688–1702.
34. Peng, K., S. Vucetic, …, Z. Obradovic. 2005. Optimizing long intrinsic disorder predictors with protein evolutionary information. *J. Bioinform. Comput. Biol.* 3:35–60.
35. Dunker, A. K., J. D. Lawson, …, Z. Obradovic. 2001. Intrinsically disordered protein. *J. Mol. Graph. Model.* 19:26–59.
36. Peng, K., P. Radivojac, …, Z. Obradovic. 2006. Length-dependent prediction of protein intrinsic disorder. *BMC Bioinformatics.* 7:208.
37. Tusnády, G. E., Z. Dosztányi, and I. Simon. 2005. PDB_TM: selection and membrane localization of transmembrane proteins in the protein data bank. *Nucleic Acids Res.* 33 (Database issue):D275–D278.
38. Deng, X., J. Eickholt, and J. Cheng. 2012. A comprehensive overview of computational protein disorder prediction methods. *Mol. Biosyst.* 8:114–121.

39. Monastyrskyy, B., K. Fidelis, …, A. Kryshtafovych. 2011. Evaluation of disorder predictions in CASP9. *Proteins.* 79 (*Suppl 10*):107–118.

40. Romero, P., Z. Obradovic, …, A. K. Dunker. 2001. Sequence complexity of disordered protein. *Proteins.* 42:38–48.

41. Linding, R., L. J. Jensen, …, R. B. Russell. 2003. Protein disorder prediction: implications for structural proteomics. *Structure.* 11:1453–1459.

42. Ward, J. J., J. S. Sodhi, …, D. T. Jones. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* 337:635–645.

43. Walsh, I., A. J. Martin, …, S. C. Tosatto. 2012. ESPRITZ: accurate and fast prediction of protein disorder. *Bioinformatics.* 28:503–509.

44. Prilusky, J., C. E. Felder, …, J. L. Sussman. 2005. FOLDINDEX: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics.* 21:3435–3438.

45. Linding, R., R. B. Russell, …, T. J. Gibson. 2003. GLOBPLOT: exploring protein sequences for globularity and disorder. *Nucleic Acids Res.* 31:3701–3708.

46. Dosztányi, Z., V. Csizmok, …, I. Simon. 2005. IUPRED: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics.* 21:3433–3434.

47. Deng, X., J. Eickholt, and J. Cheng. 2009. PREDISORDER: ab initio sequence-based prediction of protein disordered regions. *BMC Bioinformatics.* 10:436.

48. Yang, Z. R., R. Thomson, …, R. M. Esnouf. 2005. RONN: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. *Bioinformatics.* 21:3369–3376.

49. Zhang, T., E. Faraggi, …, Y. Zhou. 2012. SPINE-D: accurate prediction of short and long disordered regions by a single neural-network based method. *J. Biomol. Struct. Dyn.* 29:799–813.

50. Obradovic, Z., K. Peng, …, A. K. Dunker. 2003. Predicting intrinsic disorder from amino acid sequence. *Proteins.* 53 (*Suppl 6*):566–572.

51. Matthews, B. W. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta.* 405:442–451.

52. Noivirt-Brik, O., J. Prilusky, and J. L. Sussman. 2009. Assessment of disorder predictions in CASP8. *Proteins.* 77 (*Suppl 9*):210–216.

53. Bordoli, L., F. Kiefer, and T. Schwede. 2007. Assessment of disorder predictions in CASP7. *Proteins.* 69 (*Suppl 8*):129–136.

54. Jin, Y., and R. L. Dunbrack, Jr. 2005. Assessment of disorder predictions in CASP6. *Proteins.* 61 (*Suppl 7*):167–175.

55. Carpenter, J., and J. Bithell. 2000. Bootstrap confidence intervals: when, which, what? A practical guide for medical statisticians. *Stat. Med.* 19:1141–1164.

56. Hanley, J. A., and B. J. McNeil. 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology.* 143:29–36.

57. Eyal, E., S. Gerzon, …, V. Sobolev. 2005. The limit of accuracy of protein modeling: influence of crystal packing on protein structure. *J. Mol. Biol.* 351:431–442.

58. White, S. H., and W. C. Wimley. 1999. Membrane protein folding and stability: physical principles. *Annu. Rev. Biophys. Biomol. Struct.* 28:319–365.

59. Finger, C., T. Volkmer, …, D. Schneider. 2006. The stability of transmembrane helix interactions measured in a biological membrane. *J. Mol. Biol.* 358:1221–1228.